
# TripAdvisor European Restaurants

**By**

| | | | |
|---|---|---|---|
| **Mr.** | **Wasin** | **Heesawat** | **6288077** |
| **Ms.** | **Intr-orn** | **Lertsupakitsin** | **6288089** |
| **Mr.** | **KrissanapongPalakham** | **Palakham** | **6288102** |
| **Mr.** | **Pongsakorn** | **Piboonpongpun** | **6288107** |

**Presented to**

**Asst. Prof. Dr. Srisupa Palakvangsa Na Ayudhya**

**A Report Submitted in Partial Fulfillment of**
**the Requirements for**
**ITCS495 Special Topics in Database and Intelligent Systems**

**Faculty of Information and Communication Technology**
**Mahidol University**
**Semester 1/2022**

# Table of Contents

## 1. Business Domain

A company named TripAdvisor provides online travel research to assist consumers in creating and taking the best possible vacation. Through its flagship TripAdvisor brand, TripAdvisor's travel research platform compiles user reviews and comments about destinations, lodging (including hotels, B&Bs, specialty homes, and vacation rentals), restaurants, and activities across the world. Each month, TripAdvisor receives about 460 million unique visitors. There are more than 760 million traveler reviews on this internet site. There are reviews on TripAdvisor for almost 7 million restaurants and hotels. The TripAdvisor name is recognized in 49 markets throughout the world. [1]

The European restaurants on TripAdvisor are the subject of our chosen dataset, which is relevant to the food and travel industry. We chose this because of its relevance because food is one of the most essential elements for survival and progress in our lives. There are many restaurants in the Europe region, and they could all be examined for all attributes to produce a summary of all restaurants, determine which one will become famous, and promote it to clients or visitors through their travel agency.

Therefore, this became important to understand whether dashboards, graphs, charts, or any other visualizations may help businesses grow, promote international travel, and promote sales at individual restaurants. In order to promote international travel for Thai citizens who are interested in traveling or who want to visit overseas, this might also be integrated into Thailand for greater benefit by being recommended through online social platforms like Instagram, Facebook, and Twitter.

## 2. Target Users

In this project, we classified the target users into two groups, namely businesses and customers. These are the descriptions:

1) **Company/Business/Organization –** Whether it's a conventional or contemporary restaurant, our focus is on the travel business as well as the company and restaurant. The analysis could assist the relevant business figure out how to increase sales and what aspects of consumer feedback, both positive and negative, need to be

improved. Additionally, a number of organizations in Europe as well as in other areas could make use of this data to promote tourism businesses, including guiding tours, restaurants, and airport organizations.

2) **Individual Customer –** There was no denying that the analysis would genuinely help a number of clients. Particularly, a group of customers who enjoy eating, traveling and following food bloggers. These customers would use our analysis in a variety of ways, including decision-making, review, and recommendation. People who enjoy eating and traveling would look over the analysis before selecting a restaurant. As a result, clients and bloggers may access our study through a number of postings and blogs.

## 3. Objectives

There are several purposes that could be analyzed for the decision-making of all users. The purpose of this project includes:

- To make decision-thinking choosing a restaurant for customers
- To develop the restaurant to be able to own weaknesses or things that need to be improved.
- To promote international traveling for customers who lived in both Thailand and other countries

Regarding the type of target users, we provided the following visualization to suit them. Each visualization could be assisted the target users in order to receive more information. The visualization can be answered the question as follows:

- What are the top 10 restaurants with the highest rating in each country?
- What are the hours for all restaurants in each country and what are the opening and closing hours for each restaurant?
- How do you know if a restaurant is rated well?
- What kind of vegetarian restaurants are there and how good is each one?
- How do I know the location of the restaurant?
- Where are family-friendly restaurants, and if I have older people, where can I take care of them?

- Comparison of restaurants including countries in each section

In response to the aforementioned list of questions. The target customer's requirements still have many questions that can be answered by the analysis. Take an event as an example. The Wakanda family decided to visit Europe and settled in Italy. However, it is challenging to find a restaurant that is wheelchair accessible because they have elderly people to look after. So, we have a visualization of finding the best restaurants in Italy with cart access making it easy to find restaurants for this family.

## 4. Data Description

We provided the data description from the TripAdvisor statistics for European eateries that we discovered on Kaggle. This dataset includes a CSV file with 42 columns and around 1.9 million eateries. The following description:

1) **restaurant_link –** The unique restaurant's TripAdvisor link might serve as the dataet's primary key as the string for locating other data. There are all valid 1,083,397 values in this data, such as “g10001637-d10002227”.
2) **restaurant_name –** The name of the eatery listed on the TripAdvisor website for which iterative values might be offered. 840,914 valid distinct values, such as “Le-Saint Jouvent”, are present as strings in this data.
3) **original _location –** On the TripAdvisor website, this information provided an estimated location. There are roughly 66,000 different values present as a list of strings in this data, such as "["Europe", "France", "Ile-de-France", and "Paris"].
4) **Country –** These details offered the name of the country where the restaurant was located after retrieving its original location. As a string, there are 24 distinct countries, with Italy making up 21% of those.
5) **region –** These details offered the name of the region where eateries can be found and were collected from the original location. There are 250 distinct localities, with Lombardy accounting for 3% of them. The valid values would only have 1,030,000 values since 50,300 areas, or 5% of those locations, have missing values.
6) **province –** The province of the restaurants that were retrieved from the original location was provided by this data. There are 1,333 distinct locales, with the

province of Barcelona is the most prevalent. However, 31% or 341,000 of the missing values were located, making only 743,000 provinces valid.

7) **city –** The city where the eateries from the original location were located. There are 43,500 different cities, with Paris being the most prevalent. The data is only valid for 683,000 cities because there are around 401,000 missing values or 37% of all cities.
8) **address –** The location of the restaurants that were listed on the TripAdvisor website's map. There are 1,034,685 distinct restaurant addresses that contain every possible value, such as "Greece." This information is displayed as string data.
9) **latitude –** The restaurant's latitude, which was given as a latitude coordinate. Only 1,070,000 values are valid since there are 15% or 15,800 missing values. This data has a mean of 46.6, a standard deviation of 5.88, a minimum value of 27.2, and a maximum value of 69.9, according to statistics.
10) **longitude –** By using longitude coordinates, this data showed the restaurant's longitude value. Additionally, there are around 1%, or 15,800 seems, of missing values for latitude. According to statistics, this data has a mean of 5.84, a standard deviation of 8.64, a minimum value of -71.2, and a maximum value of 33.4.
11) **claimed –** This data, which is displayed as a string, represents the restaurants that TripAdvisor has claimed or not claimed. The most popular restaurant, however, is unclaimed and only has 1,842 missing values.
12) **awarded –** This data is representing an award name that could be belonged to the restaurants. However, there are some variables that are missing, which may indicate that the restaurant did not receive any incentives. 820,000 data, or almost 76% of the total, are missing. Additionally, 917 has distinctive qualities, and the most typical recognition is the Travelers' Choice, Certificate of Excellence 2020.
13) **popularity_detailed –** The data displayed as a string in this data represented the restaurant's ranking in various categories. This data is valid and usable, however, 95,000 data points, or 9% of the total, are missing from it.
14) **popularity_generic –** This data displays the ranking of restaurants as strings in a particular area, with each restaurant's rating number in relation to others in the same

industry. There are only 982,000 valid values in this set of data because 9% or 97,800 of the 982,000 unique values were missing.

15) **top_tags –** This data displays the tag as a keyword-searchable string that includes terms like Japanese, Sushi, and Asian, among others. This data comprises 973,000 valid records and 111,000 missing records. The most prevalent tag in this set of 40,000 unique data is "Mid-range, French."
16) **price _level –** According to three distinct types of currency signatures; € for low prices, €€-€€€ for medium prices, and €€€€ for high prices; the data shows the pricing range for each establishment as a string. For 50% of those statistics, the average restaurant price is medium price. Additionally, there are around 277,000 missing values or 26% of the total data, so only 806,000 entries are present in the legitimate data.
17) **price_range –** Along with price level, there is also price range, which displays the range of prices in a currency as a string, like €10-€30. Approximately 779,000 values, or 72%, of the data, are missing, making it valid for just 304,000 values, or 28%, and 7298 unique values.
18) **meals -** The data depicts the many meals of each restaurant that are displayed as strings, such as lunch, brunch, and dinner. Only 635,000 of the 745 unique data points are legitimate, making up 48% of the total data. There are 448,000 missing values.
19) **cuisines –** For each restaurant, the types of food are displayed here as strings, such as Italian. It is valid for around 914,000 values because there are 97,700 unique values and 169,000 (16% of the total) missing values. This statistic demonstrates that Italian restaurants are the most prevalent in the region of Europe.
20) **special_diets –** This data displays different kinds of special diets in strict terms like vegetarian-friendly. About 743,000 or 69% of the missing values have 68 unique values, which reflect multiple establishments that are not diet eateries. As a result, it was determined to be valid for 340,000 diet restaurant values.
21) **features –** This data displays each restaurant's unique systems or features, such as reservations, seating, accessibility, serving alcohol, and table service. 56,500 different values, or 766,00 missing values, represent various restaurants that lack

attributes or are not specified in the data set. Therefore, there are only 317,000 viable values in all.

22) **vegetarian_friendly –** This data demonstrates whether or not this restaurant is vegetarian-friendly. This information displays Boolean type (true or false). 30% of the answers, or 324,000, are accurate, while the remaining answer is wrong.
23) **vegan_options** – This Boolean response determines whether the restaurant provides vegan options or not. Almost all of the data has 13% or 137,000 "true" responses and 87% or 947,000 "false" answers, both of which are genuine.
24) **gluten_free –** This data shows the restaurant's gluten-free menu choices. This feature is available at certain restaurants but not all. There are 1,080,000 different values with Boolean types, with true representing 123,000 or 11% of the values and false representing 960,000 or 89% of the values. This means that many restaurants do not offer this choice.
25) **original_open_hours –** This data displays the typical opening times for each restaurant on TripAdvisor. We could only use 594,000 unique values because the string's map to the missing data contains 490,000 unique values, or 45% of those, leaving 238,000 unique values. Data such as “Mon”: [“00:00-23:59”] represents the length of time the restaurant was open and closed on each day.
26) **open_days_per_week –** This data reveals the total number of days that each restaurant was open during a given week as integer. For example, some restaurants were open seven days a week, while others were only open six. This data only has 594,000 valid values since there are around 490,000 missing values, or 45%, which is a high null value. For statistical purposes, the days a restaurant was open were 1 to 7 days, where 1 was the shortest day and 7 was the longest. The mean of this data was 6.33, and the standard deviation was 0.97.
27) **open_hours_per_week –** This data displays each restaurant's total weekly hours of operation as a float. We could only use 594,00 values because there were too many missing values in original open hours (490,000 or 45% of the data were missing). 0 is the minimum value, 39 is the 25% quantile, 58.5 is the 50% quantile, 81.5 is the 75% quantile, and 168 is the maximum value in terms of statistics. This data has a mean of 62 and a standard deviation of 30.5.

28) **working_shift_per_week –** This data displays, as an integer, the number of working shifts per week that were derived from the original open hours. We could only use 594,000 values because 490,000 or 45% of the data's missing values are present. According to the statistics, 1 represents the smallest value, 6 represents the 25%, 50%, and 75% quantiles, and 15 represents the maximum value. The data's mean is 7.63 and its standard deviation is 2.55.

29) **avg_rating –** This information displays the typical reviews of each restaurant as a floating star that represents the contentment of the patrons. We might use the 987,000 valid values in this set of data since there are 96,600 invalid values or 9% of the total. According to the statistics, the standard deviation is 0.71 and the mean rating is 4.04. A minimum value is 1, a 25% quantile is 3.5, a 50% quantile is 4, a 75% quantile is 4.5, and a maximum value is 5

30) **totals_reviews_count** – The total number of reviews for each restaurant is displayed in this data as an integer. While there are 1,030,000 valid values or 95% of them, only 52,200 are missing. The statistics show that 0 is the least value and 52,400 is the largest value. The mean and standard deviation for the entire sample of reviews is 103 and 267, respectively.

31) **default_language** – This data demonstrates the default language of the website. There are roughly 95,200 or 9% and 2 unique values, making up about 988,000 or 91% of the permissible values, including English.

32) **review_count_in_default_language** -The total amount of review counts in the default language are shown in this data. 98% of the values in this data, or 988,000, are valid; 9%, or around 95,200, are missing. This data has a mean of 44.6 and a standard deviation of 149 for statistical purposes. The maximum value is 15,200, with a minimum of 1, a 25% quantile of 2, a 50% quantile of 7, a 75% quantile of 26, and a 75% quantile of 1.

33) **excellent** – The number of 5-star reviews in the default language is displayed as an integer in this data. It has 95,200 missing records and 988,000 valid records, or 91% of the total. According to the statistics, the good review count for this data is on average 24.7, and the standard deviation is 89.9. Additionally, it has a maximum

count of 9,380 and a minimum count of 0. The quantiles of 25%, 50%, and 75% are 1, 3, and 13, respectively.

34) **very_good** – This data shows that the number of extremely positive ratings is an integer in the default language. 91% of the data—roughly 988,000—is legitimate, and 9%—roughly 95,200—is missing. The highest count is 4,090, the minimum count is 0, the quantile for 25% is 0, the quantile for 50% is 2, the quantile for 75% is 6, and the quantile for the maximum is 6. While the standard deviation is 35.5 and the mean of the average review count is 10.5.

35) **average** – This data displays the average review count as an integer for the default language. 91% of the data are valid (about 988,000), and 9% (about 95,200) are missing (about 988,000). The information shows that the standard deviation is 15.7 and that the mean average review count is 4.11. The number of counts ranges from 0 to 2,131. The quantiles for 25%, 50%, and 75% are 0, 1, and 2, respectively.

36) **poor** – The number of negative reviews is represented as an integer in this data using the default language. 95,200 or 9% of the data are missing, leaving 988,000 or 91% of the data to be legitimate. This data's standard deviation is 9.35 and it is mean is 2.36. The statistic's minimum and maximum values are 0 and 1,250, respectively. While the quantiles for 25%, 50%, and 75% are all 0, 1, and 0, respectively,

37) **terrible** – These numbers display the number of negative reviews in the selected language. It has 95,200 or 9% missing records and 988,000 or 91% valid records. According to the statistic, the quantiles for 25%, 50%, and 75% are 0, 0, and 2, respectively. The range of counts is from 0 to 1,220.

38) **food** – This data indicates how each restaurant's meal rating floats. 484,000 or 45% of the values are missing whereas 599,000 or 55% of the values are legitimate. Additionally, the data's mean is 4.1 and its standard deviation is 0.56. The quantiles for food ratings are 1, 4, and 4.5, with 1 being the lowest and 5 being the highest.

39) **service** – In this data, the service rating for each restaurant is displayed as a float. It consists of 604,000 legitimate records, or 56%, and 479,000 missing records, or 44%. The statistic demonstrates that the mean service rating for this data is 4.07, and the standard deviation is 0.58. Furthermore, it has a minimum rate of 1 and a

maximum rate of 5. The quantiles for 25%, 50%, and 75% are 4, 4, and 4.5 respectively.

40) **value** – This data indicates each restaurant's value rating as a float. About 56% of the valid value, or 603,000, and 44% of the missing value, or 481,000, are present. The statistic shows that the value rating has a mean of 3.98 and a standard deviation of 0.58. The quantiles for 25%, 50%, and 75% are respectively 3.5, 4, and 4.5. While the minimum and maximum rates are 1 and 5, respectively.

41) **atmosphere –** In this data, each restaurant's atmosphere rating value is displayed as a float. It contains roughly 822,000 or 76% and about 262,000 or 24% of the valid value. This data's standard deviation is 0.56 and it is mean is 3.93. Additionally, the 25%, 50%, and 75% quantiles are 3.5, 4, and 4.5, respectively. The minimum and maximum rates are 1 and 5, respectively.

42) **keyword** – These data display as a string the top keywords for each restaurant, such as "steak," "onion loaf," "lettuce wedge," "chateaubriand," and "tbone." Because there are 99,000 unique values and 91%, or 984,000 missing values, it is valid for 9%, or roughly 99,200 values.

## 5. Alteryx Workflow

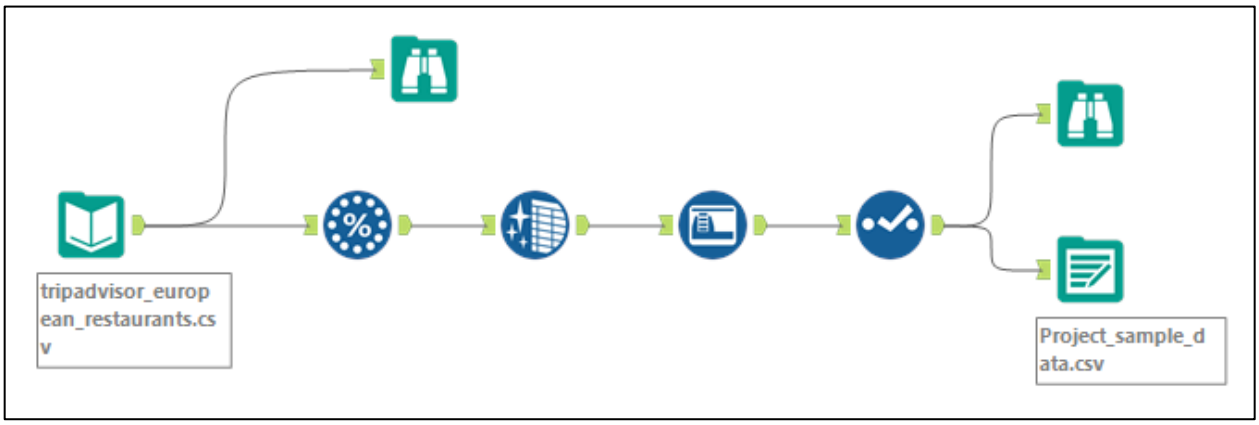



Figure 5.1: The workflow for fetching the data

This workflow is a process for fetching the original data to 500,000 data because the original data have more than one million data, so we decrease the data that we use for decreasing the time process. We started by inputting the original data which is the excel data. Next, we use a

random sample that set the data in 500,000 data, then we clean the data-by-data cleansing to check the missing data and use auto field and select to set field type and make the data have the smallest possible size. Last, we use output data to get the result into the file as a CSV file.

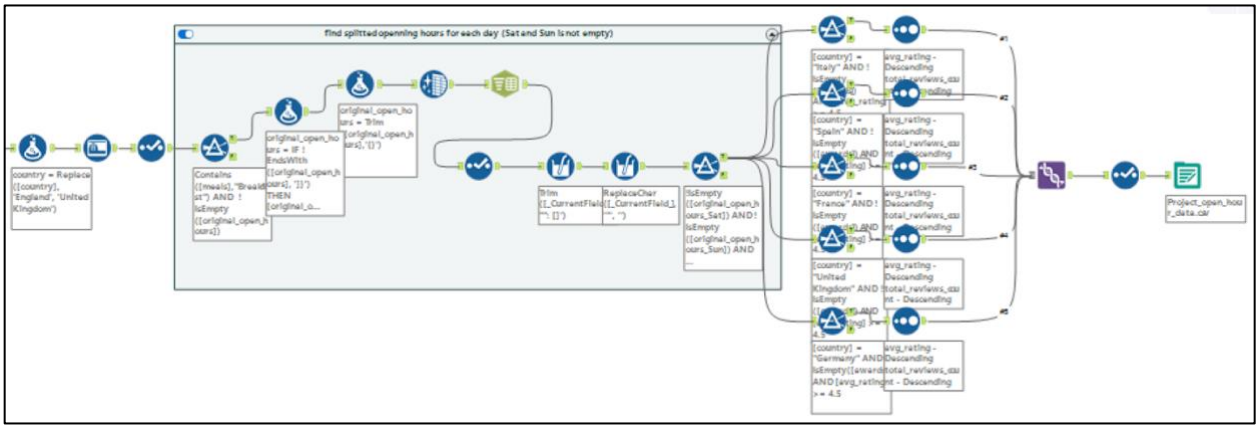



Figure 5.2: The workflow for family-restaurant flow

This workflow is a process of finding open hours for Saturday and Sunday which is a holiday on the weekend, so several families used this day to go on a vacation. We started by extracting the output from Figure 5.1, as you can see in the first container. After we extracted, the data was preliminarily cleaned by replacing “England” in column country with “United Kingdom” for matching the map in Power BI. Then, we cleaned by auto-filling to change the type of data appropriately and check what data need to be cleaned more with the select function.

After we are preliminarily cleaned, we cleaned more in order to find the opening hours for Saturday and Sunday of each restaurant. The first step is to check the data that meals contained a breakfast and that original open hours are not null. Then, we checked if the original open hour field is not ended with “]}”, it will be added by “0” because of erroneous information. Then, we trimmed all punctuation around the data to take only time. After we filtered, we cleaned by auto-filling to check the type of data and separated the original open hours into columns. After that, we change the name of all new columns with “original_open_hours_{day}” (day means Monday to Sunday respectively). When setting all new column name finish, we trim the data to remove unwanted characters that have quotation marks, colons, and square brackets. Next, we replace the remaining quotation marks that can be occurred in case of the restaurant has several ranges of opening hours, so we replaced it with an empty value. After we finished replacing them, we check

the stores that are open on Saturdays and Sundays by filtering to check the opening hours of Saturdays and Sundays are not null.

When we filtered the data on which restaurants were open on weekends, we selected the top 10 restaurants for each country starting by filtering each name of country and award required and an average rating greater than or equal to 4.5. Then, we sort the data from the average rating and total reviews count set as descending. Lastly, we use the union tool to merge the data of each country together and then select the column we want and then come out as an output file.

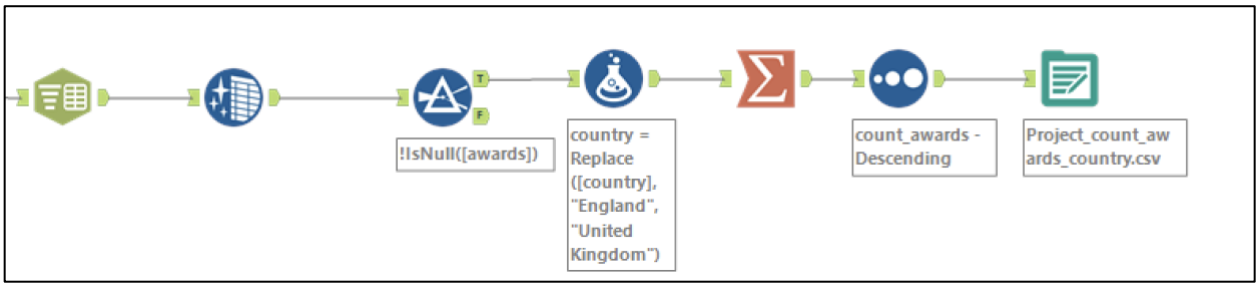



Figure 5.3: The workflow for the number of awards

This workflow is a process of summarizing awards for each restaurant. The workflow starts by extracting the output of Figure 5.1. We separated the awards of each column by dashed punctuation because some restaurants have more than one award. After we changed the data to the column, we have to clean it by using data cleansing to remove unwanted whether characters and punctuation. Then, we select only existing awards and also replace a country field from England to the United Kingdom due to it cannot find a location in the map of Power BI. Then, it was grouped by restaurant name, country, latitude, longitude, and region in order to summarize the number of awards for each restaurant. We finished it by sorting the number of awards in descending order and exporting an output for the CSV file.

## 6. Power BI Visualization

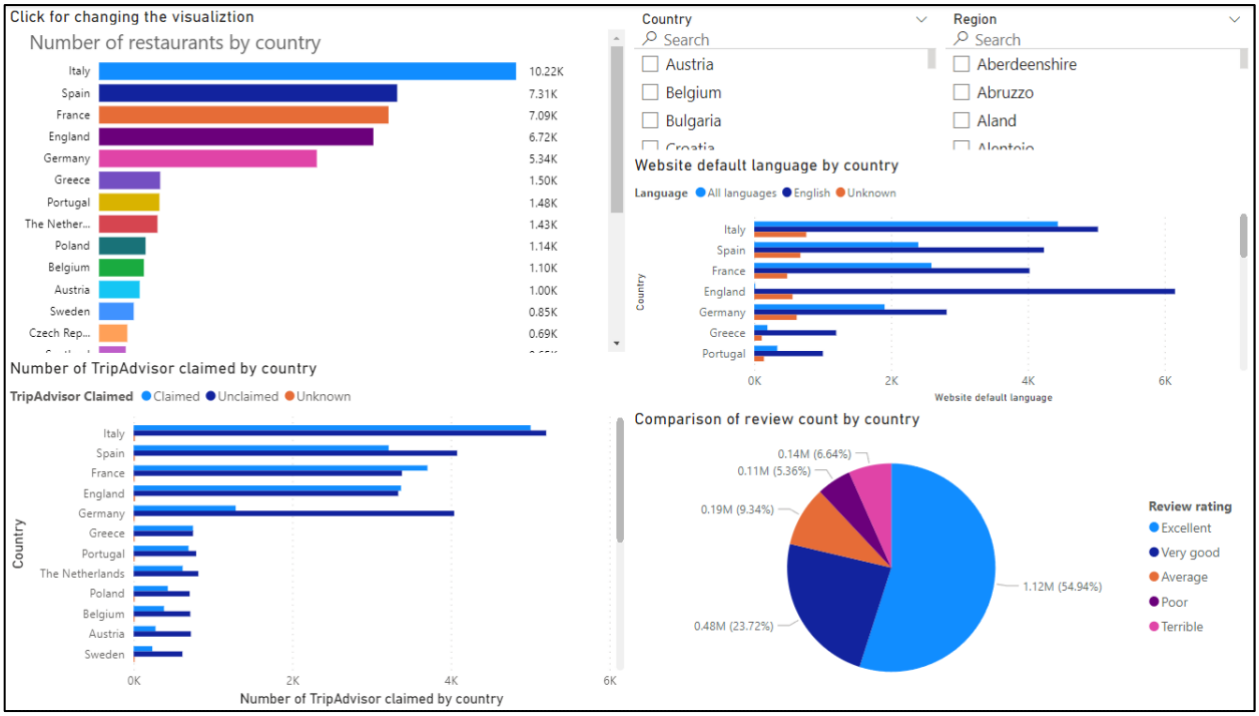



Figure 6.1: Comparison of information by country

According to Figure 6.1, we can see the dashboard provided several visualizations about the sum or count for each country. The first visualization is the number of all restaurants in each country Italy has the highest number of restaurants about 10,220 stores. Below one is the number of restaurants that are claimed by TripAdvisor in each country. Italy also has the highest number of claimed restaurants that may be occurred from the number of restaurants already. However, Italy has the unclaimed restaurant more than claimed. The next one is a comparison of the website language of restaurants in each country. Most restaurants are using the English language as a default language because English might be a general language in our world and restaurants in that country can speak the English language. The pie chart is a comparison of the review count in each country. It represents the five sections including excellent, very good, average, poor, and terrible reviews. Figure 6.1 is a comparison of review counting for all countries which is mostly an excellent review. Furthermore, we can select the filter by slicers at the top of the dashboard. The slicer contains the country selection and region selection in order to deep dives more for specific information.

However, there is some hidden visualization in Figure 6.1 located on the top-left side of the dashboard. As you can see in the first visualization, we will see the text “Click for changing the visualization” that can be changed to the other visualization. Those visualizations are consisted of:

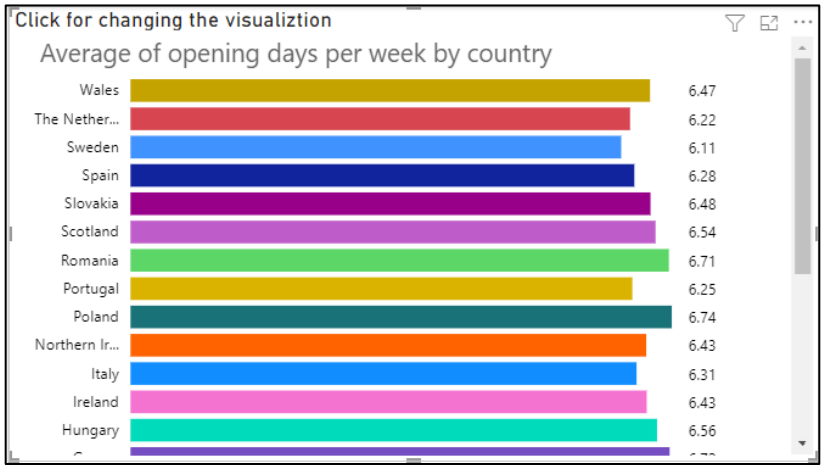



Figure 6.2: An average of opening days per week by the country

This visualization shows the average opening days per week for each country in order to know how much the different average hours are.

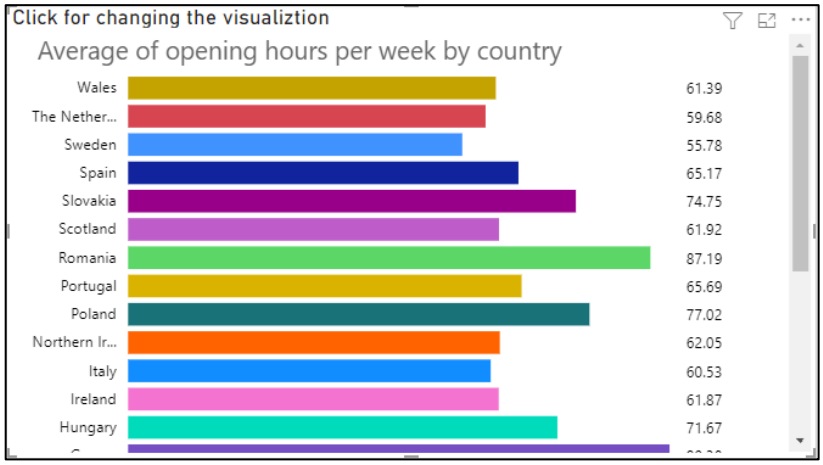



Figure 6.3: An average of opening hours per week by the country

This visualization represents the average opening hours per week for each country. This could be told about hours of operation being relatively low, indicating that the restaurant may close early and is not suitable for traveling as a kitchen.

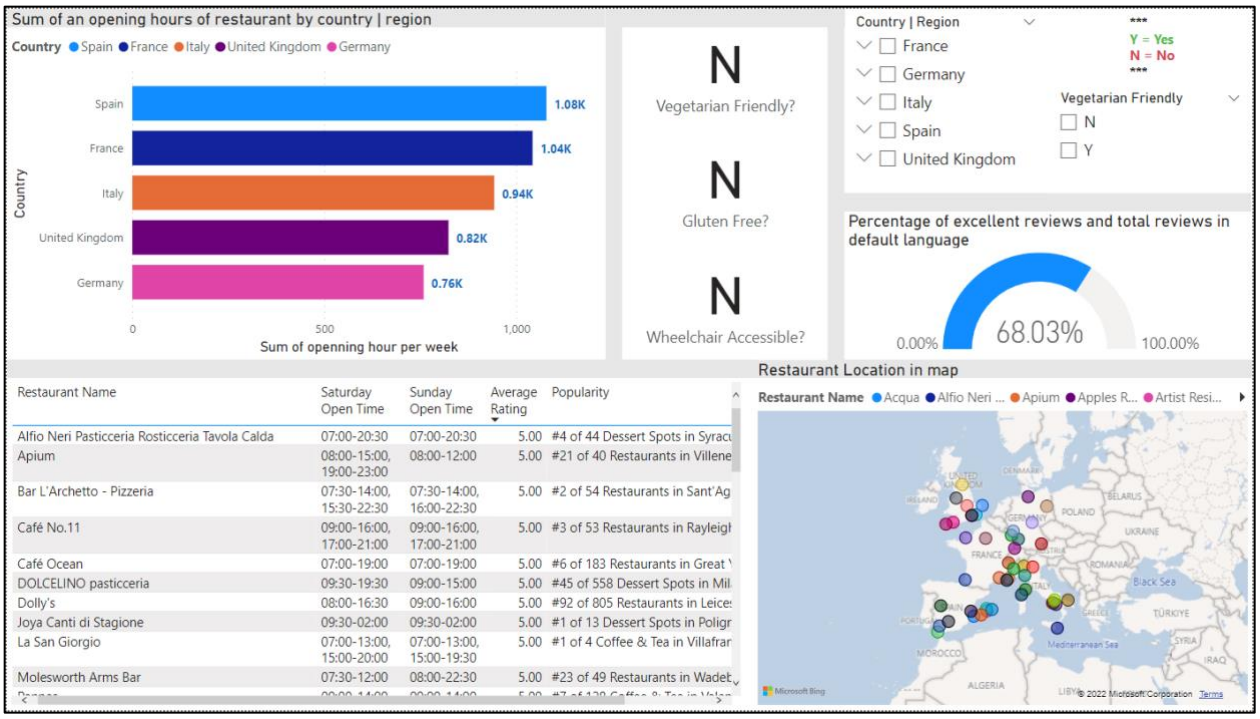
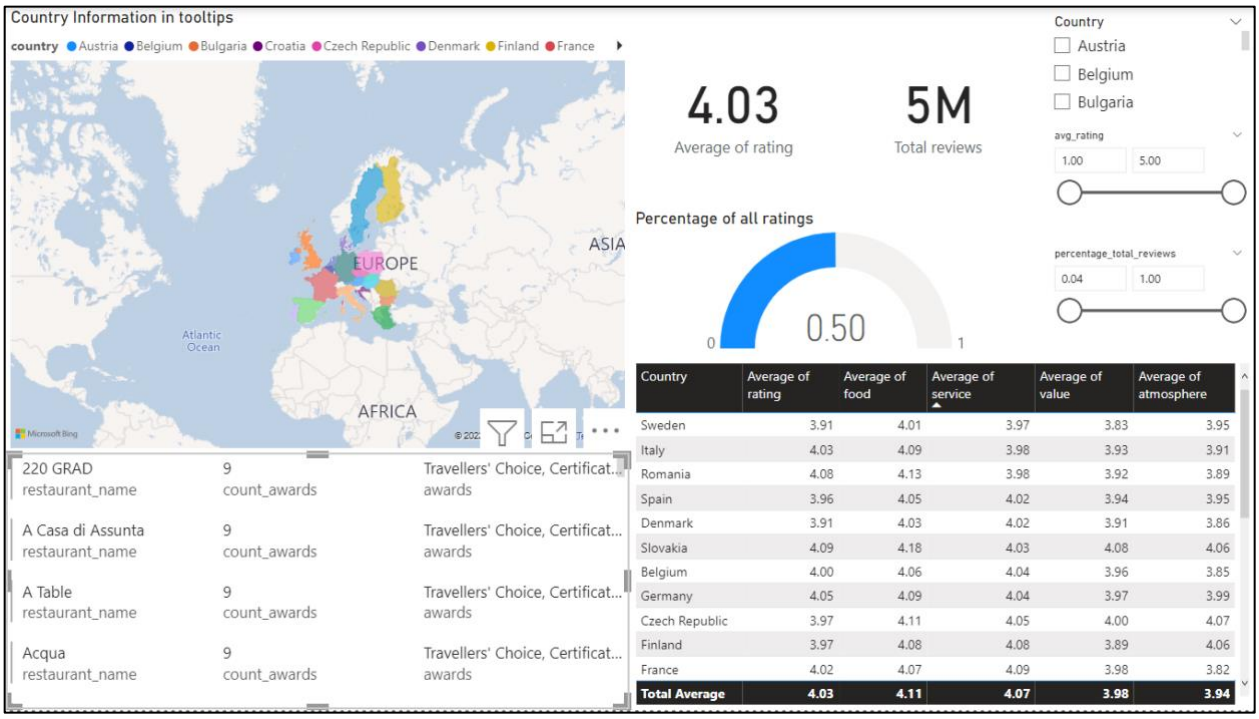



Figure 6.4: Top 50 family-friendly restaurants by the country

According to Figure 6.4, the dashboard contains the top 10 average ratings for each restaurant. However, we sum up the country that has the highest number of restaurants. The first one is the bar chart about the sum of opening hours of restaurants in each country Spain has the highest number of hours about 1,080 hours. Moreover, we can do deep dives by clicking on the chart of each country to find insights in terms of region. At the bottom of the dashboard, we provide a table presenting a restaurant name, open hours on Saturday and Sunday, the Average rating of the restaurants, and general popularity in the country. This visualization also provides a restaurant's location represented as a map that can tell the users where restaurants are located. Moreover, if we select once a restaurant it will show more insights. There are several cards at the top of the dashboard including vegetarian-friendly, gluten-free, and wheelchair accessible. This visualization represents whether the chosen restaurant is contained the features or not. If a card shows "Y", it means the restaurant has the feature; but if a card shows "N", it means the restaurant does not have those features. Another interesting visualization is a gauge bar representing a percentage of an excellent review and total reviews in terms of the default language. In addition, there are slicers including country-to-region selection and vegetarian-friendly options in order to specify more insights into what we want.




| Country | Average of rating | Average of food | Average of service | Average of value | Average of atmosphere |
|---|---|---|---|---|---|
| Sweden | 3.91 | 4.01 | 3.97 | 3.83 | 3.95 |
| Italy | 4.03 | 4.09 | 3.98 | 3.93 | 3.91 |
| Romania | 4.08 | 4.13 | 3.98 | 3.92 | 3.89 |
| Spain | 3.96 | 4.05 | 4.02 | 3.94 | 3.95 |
| Denmark | 3.91 | 4.03 | 4.02 | 3.91 | 3.86 |
| Slovakia | 4.09 | 4.18 | 4.03 | 4.08 | 4.06 |
| Belgium | 4.00 | 4.06 | 4.04 | 3.96 | 3.85 |
| Germany | 4.05 | 4.09 | 4.04 | 3.97 | 3.99 |
| Czech Republic | 3.97 | 4.11 | 4.05 | 4.00 | 4.07 |
| Finland | 3.97 | 4.08 | 4.08 | 3.89 | 4.06 |
| France | 4.02 | 4.07 | 4.09 | 3.98 | 3.82 |
| **Total Average** | **4.03** | **4.11** | **4.07** | **3.98** | **3.94** |

Figure 6.5: Rating by the country

According to Figure 6.5, the dashboard shows ratings of restaurants in each country. We separated into two sides including a map visualization and an information visualization. The left side is the map related to the country of a data source. We will see the highlighted country that represented those countries are contained in the data source. If we position the clicker on a country, it will show the tooltip details. The right side shows the information about ratings represented as whether a card, gauge, or table. The main table represents information on several kinds of averages including the average rating, the average food, the average service, the average value, and the average atmosphere. All kinds of averages are classified for each country. The gauge bar represents the percentage of all kinds of ratings which is separated into five categories, thus it would be divided by 25 because each review can be the highest value of only 5 points. Overall, will be 50% of all as a default value. A couple of cards represent the average rating and total of reviews in order to make users try to look in a wide perspective. There also have multiple cards representing the number of awards for each restaurant. It represents rating and award looks appropriate or not. In addition, there are three slicers for filtering including a country selection, a range of the average rating, a range of total reviews in terms of percentage, and a range of the number of awards.

## 7. References

[1] TripAdvisor, "Investor FAQs," [Online]. Available: https://ir.tripadvisor.com/investor-faqs. [Accessed 15 November 2022].